論文内容要旨（和文）

平成 18 年度入学　大学院博士後期課程
システム情報工学専攻　生体数理情報学講座
学生番号　06522305
氏　　名　Varga István　㊞

論文題目：Machine translation resource development for low-resourced language pairs

近年における機械翻訳の発展には二つの理由が考えられる。一つ目は文法に基づく方法や統計的情報を利用する、精度の高い翻訳方法であり、二つ目は単一言語の知識表現から二言語コーパスまでの幅広い機械翻訳資源である。しかし、理論的にはほとんどの翻訳方法がすべての言語ペアに適用できるが、多くの言語ペアで最も基本的な翻訳資源が不足しているため、実際には実現不可能である。

本論では翻訳資源が少ない言語ペア、また使用頻度が少ない言語ペアの機械翻訳の可能性を探究する。この問題に対して二つのアプローチが考えられる。既存の翻訳システムの再利用と、機械翻訳資源の自動生成である。

対象とする言語ペアには使える人的資材も少ない。そこで、提案する解決方法が多くの言語ペアに適応できることが重要になる。そのためにはローコストとロバストな解決方法が必要である。以下にその方法を述べる。

本論の前半ではトランジティブ機械翻訳、つまり中間言語を介した翻訳方法を研究する。成功条件として、対象となる言語と中間言語の特性、また構成する機械翻訳システムの精度があげられる。

本論の後半では機械翻訳のための二言語翻訳資源を自動生成する新しい方法を提案する。自然言語を記述するには「内容」と「構造」の両者が不可欠である。「内容」は言語に使われている語彙、「構造」は語彙を支配する構文や文法規則に現れる。二言語の設定では二言語辞書とトランスファー規則が内容と構造の最も基本的なものだと考えられる。

二言語辞書の自動生成には中間言語への辞書とワードネットを利用する。第一ステップとして利用する辞書において原言語と目標言語のリンクが翻訳候補として生成される。第二ステップでは語彙データベースから得られた情報で交換時に入った雑音を消去する。

トランスファー規則の作成には小さなパラレルコーパスを利用する。構文解析したコーパスから最も頻度が高い文法規則や文パターンに関するトランスファー規則を自動的に推定する。

本論で提案した資源生成方法を、機械翻訳資源が少ない言語ペアの一例としてハンガリー語・日本語で例証している。また、以上の方法をフリーツールとしてインプリメントし、自由に使用できるようにしている。

論文内容要旨（英文）

平成 18 年度入学　大学院博士後期課程
システム情報工学専攻　生体数理情報学講座
学生番号　06522305
氏　　名　Varga István　㊞

論文題目：Machine translation resource development for low-resourced language pairs

The rapid evolution of machine translation in recent decades can be attributed to two main factors: highly efficient machine translation methods, varying from grammatical to statistical models and development of translation resources, ranging from monolingual knowledge representations to bilingual corpora. However, while theoretically most translation methods can be implemented with all languages, practically this is not always achievable, since numerous language pairs lack even the most basic translation knowledge.

This thesis explores the prospects of machine translation between low resourced languages. We can think of two basic methods in dealing with this problem: (1) re-using already existing machine translation systems; (2) generate the necessary translation resources. Because of the low-resourced characteristic of the chosen languages, it has to be low-cost and robust for assimilation purposes, regardless of the solution method, in order to be implementable with any language pair.

First, this thesis challenges the problem by investigating the possibility of transitive machine translation, with the introduction of a third, intermediate language between the two initial languages. The main purpose is to determine whether the characteristic of the constituent languages or the accuracy of the constituent systems plays a bigger role in a successful transitive machine translation system.

Next this thesis deals with the problem by proposing new bilingual translation resources. With the lexicon and the grammatical rules being the elements that best describe the basics of the *content* and *structure* of a language, a method for each in a bilingual environment is proposed. The bilingual dictionary generating model uses an intermediate language and WordNet to generate a new bilingual dictionary. The dictionary is used to connect entries in the source and target languages that are possible translations of each other, while with the lexical database the erroneous links are eliminated. The grammatical rule/sentence pattern generating model uses a small parallel corpus of the two languages to generate the most frequent basic grammatical rules and sentence patterns.

The proposed methods are exemplified with the Japanese-Hungarian language pair. Both translation resource generating methods are also available as an implemented ready-to-use tool.

12pt シングルスペース 300 語程度）

別　紙

専 攻 名	システム情報工学専攻	氏　名	VARGA István

学位論文の審査結果の要旨

機械翻訳においては、翻訳対象となる２つの言語の間の翻訳辞書や文法情報などの翻訳資源が必要不可欠である。本論文は、翻訳資源が少ない言語ペア、また使用頻度が少ない言語ペアに対する機械翻訳の可能性を探究したものである。例として、ハンガリー語と日本語とのペアを用いているが、手法そのものは、どのような言語ペアにも適用できるものである。

第１章は、機械翻訳の代表的な手法、すなわち、規則に基づく手法、例文に基づく手法、統計的手法の３つを紹介し、機械翻訳システムと言語の問題に言及している。また、論文のアウトラインについて述べている。

第２章は、研究の目的である、機械翻訳資源の生成、特に二言語辞書の生成と、翻訳用の文法規則の生成について述べている。

第３章は、中間言語を介した翻訳方法(transitive machine translation)について述べている。事例として、ヨーロッパの諸言語間での機械翻訳の評価を言語ペアごとに比較している。また、ハンガリー語と日本語との間で英語を介した実験を行い、結果として生じる誤りをどのように防いで、翻訳精度を向上させるかについて言及している。

第４章は、二言語辞書の具体的な生成方法を述べている。これまでの単純な中間言語を介する方式に対して、言語に使われている語彙を「内容」として記述し、語彙を支配する構文や文法規則を「構造」としてとらえている。この章では「内容」の部分を扱う。辞書の語彙的な部分である二言語辞書を、中間言語とワードネットを利用して精度を高める方法が、２つのステップ、すなわち、原言語と目標言語のリンク、データベースの情報を用いた雑音除去を用いて、具体的に述べられている。

第５章は、「構造」の部分の作成である。二言語間の翻訳を行うには、トランスファー規則が必要である。ここでは、小さなパラレルコーパスを利用して、構文解析したコーパスから、最も頻度の高い文法規則や文パターンに関するトランスファー規則を自動的に推定している。

第６章は、結論で、本論文の内容をまとめている。

このように、本論文は、機械翻訳資源の少ない言語ペアで、どのように効率的で頑健な機械翻訳資源が構築できるかを、ハンガリー語と日本語との間で例証したもので、同じ手法を、資源の少ないペアに適用できる画期的な研究成果である。

本論文の内容の一部は、国際的な学術雑誌に１編掲載され、機械翻訳や自然言語処理関連の国際会議に６編の発表を行っている。

以上のように、本研究は、学術的および工学的に価値があり、今後も応用が期待されるすぐれた知見を多数含んでおり、博士（工学）の学位論文として合格と判定する。

最終試験の結果の要旨

本学の規定に従い、本論文および関連分野に関して、口頭による最終試験を行った。その結果、申請者は、関連分野を含めて、広い分野における基礎学力を有し、既存の関連研究や関連論文の内容にもよく通じていることが分かった。また、研究についても、その手法や態度において、十分な能力を有していることが確認された。

以上のことから、申請者は博士に相当する十分な学力を有しており、博士（工学）の学位授与に関する最終試験に対して合格であると判定する。